Mathematics

https://www.aimspress.com/journal/Math

AIMS Mathematics, 11(3): 5340–5364.
DOI: 10.3934/math.2026220
Received: 03 December 2025
Revised: 20 January 2026
Accepted: 02 February 2026
Published: 02 March 2026

Research article

Accurate saddlepoint approximations for clustered rank tests under randomized block urn design

Haidy A. Newer[1,*]**and Bader S. Alanazi**[2]

[1] Department of Mathematics, Faculty of Education, Ain Shams University, Cairo 11511, Egypt

[2] Department of Mathematics, College of Science, Northern Border University, Arar 73222, Saudi Arabia

* **Correspondence:** Email: haidynewer@edu.asu.edu.eg.

Abstract: Randomized block urn design is a clinical trial design that is increasingly being used as a means of treatment allocation to balance the treatment allocation. With repeated measures or multi-centre trials though, the corresponding complex structures of dependency invalidate the standard asymptotic approximations, even in small to moderate samples. In this paper, a high-accuracy saddlepoint approximation model of a linear rank test is established with the dual conditions of cluster sampling and adaptive urn randomization. We obtained the joint cumulant generating function of the test statistic conditional in the realized block allocation counts, and the accurate calculation of mid-p-values that takes into consideration the discreteness of rank scores. We have a wide variety of classes of score functions, such as the log-rank, GehanWilcoxon and MannWhitney statistics. Wide-scale simulation experiments showed that the suggested saddlepoint algorithm is much more effective at controlling Type-I error rates and preserving nominal coverage probabilities of confidence intervals, and is comparable to computationally-intensive permutation benchmarks even with strong intra-cluster correlation. Its approach was demonstrated by applying it to oncology and ophthalmology trial data demonstrating its soundness in finite-sample cases.

Keywords: saddlepoint approximation; clustered data; linear rank tests; randomized block urn design; log-rank test; Gehan–Wilcoxon test; Mann–Whitney test; Prentice–Wilcoxon test; confidence intervals
Mathematics Subject Classification: 62F12, 62G10, 62F03, 62J12, 62N01, 62C12, 62G09, 60K35

1. Introduction

Clinical trials are often faced with a situation of clustered data, where the data is clustered in a meaningful manner, e.g. several measurements of one patient or patients in hospitals. Such structures in itself cause correlation, which invalidates the independence assumptions of most standard statistical techniques and makes valid inference difficult. In order to address these problems, it is important to use stringent randomization plans so that the treatment might be distributed equally and the bias is reduced to the minimal level.

The block urn design (BUD) by Zhao and Weng [40] is one of such advanced schemes. The BUD is customized in sequential clinical trials in two or more treatments and combines the concepts of the classical block designs with the adaptive dynamics of the urn model. The design balances the treatment groups dynamically by using a system of active and inactive urns, and the design remains random. The mechanism is especially favorable in trials that have clustering data that subunits are assigned to various treatments.

Operationally, in the BUD, the treatments are allotted by the active urn by drawing balls. After every draw, the composition of the urn is adjusted to eliminate the existing imbalances, which minimizes selection bias and enhances homogenous allocation. This adaptive property is particularly desirable in cases where the subjects come in in a serial fashion and must be assigned immediately, or when the overall sample size and subgroup sizes are not known beforehand such as is often a fact in clinical studies.

To illustrate the underlying mechanism, consider the generalized Friedman urn model [7], which serves as the foundation for the BUD. An urn initially contains γ white and γ red balls. When a ball is drawn at random and replaced, α balls of the opposite color are added. This process iterates throughout the trial. Drawing a white ball assigns a subject to treatment A, while a red ball assigns them to treatment B. In the special case where $\gamma = 0$ and $\alpha > 0$, the first assignment occurs with equal probability (0.5). If, at a given stage, i subjects have been assigned to treatment A and j to treatment B, the probability that the next subject is assigned to treatment A becomes $j/(i + j)$. It is a very straightforward but useful mechanism that has been proven to estimate optimal random allocation designs of predetermined group sizes [36].

At the limit of $\alpha = 0$, the design reduces to complete randomization indicating the flexibility of the urn framework. The degree of balance achieved is quantified by the absolute difference in treatment assignments at stage i, denoted by D_i. This difference evolves as a stochastic process with transition probabilities governed by γ, α, the current imbalance d, and the number of assignments i [37]:

$$p_i = p(i,d) = \begin{cases} \mathrm{P}(D_{i+1} = d - 1|D_i = d) = \frac{1}{2} + \frac{\alpha d}{2(2\gamma+\alpha i)}, & d = 1, 2, \ldots, i, \\ \mathrm{P}(D_{i+1} = d + 1|D_i = d) = \frac{1}{2} - \frac{\alpha d}{2(2\gamma+\alpha i)}, \\ \mathrm{P}(D_{i+1} = 1|D_i = 0) = 1. \end{cases} \tag{1.1}$$

These probabilities reveal that the likelihood of reducing imbalance increases with the magnitude of the current imbalance d but decreases as the trial progresses (increasing i), eventually converging to 0.5. Thus, the urn design UD(γ, α) effectively forces balance early in the trial while mimicking complete randomization as the sample size grows.

Although randomization in the BUD is on the individual level, statistical analysis should strictly consider the intra-cluster correlation of the data. Dependencies permeate between and within treatment groups when correlated subunits in clusters are randomised to different arms. This correlation is very challenging to infer especially in survival analysis whereby censoring is common.

In order to compare independent survival distributions under these conditions, a number of nonparametric tests based on ranks have been developed including the Gehan-Wilcoxon test [8], the Prentice–Wilcoxon test [28] and the log-rank test [34]. The rank-based tests like Mann-Whitney U-test are strong substitutes to parametric tests since they are not dependent on the assumption of normality. The logic of the pairwise comparison underlying the Mann-Whitney statistic has been recently brought back to the limelight by the so-called win ratio method of composite endpoints (Pocock et al. [27]). This framework sees the statistic as a number of "wins" to number of "losses", which is also subject to criticism in recent literature on ties and intransitivity (Oakes [26]). But even conventional uses of these tests do not take into consideration intracluster correlations. Rosner and Grove [32] overcame this by generalizing the MannWhitney U-test to include several parameters of correlation, while Jeong and Jung [12] derived adjusted rank tests of specifically clustered survival data. Our research is based on these foundations and develops inference tools that are specially designed to address the two complexities of the randomized block urn design and clustered rank statistics.

Saddlepoint approximation is an effective method of enhancing the accuracy of p-value and confidence interval estimation especially in cases where sample sizes are small or the data distributions are complicated. The most recent extensive reviews have highlighted the flexibility of the technique and shown it to have been superior to traditional approximations in a variety of areas, such as reliability analysis (Meng et al. [16]). Whereas conventional asymptotic techniques are based on the normality assumptions that are often not valid in finite samples, saddlepoint approximations make use of the complete cumulant generating function to approximate exactly the distribution of test statistics with spectacular accuracy [4, 5, 29]. This has largely been effective in rank based tests in randomized block designs which provide accurate tail probabilities at high computational efficiency [19–21]. Outside the clinical statistics. It forms a basis of modern reliability engineering, where it has been applied successfully to quadratic functions of normal variables and reliability-based optimization with Gaussian mixture models [11, 16, 17].

The saddlepoint approximations give a solid framework of inference in the context of clinical trials with clustered data and sequential allocation as in the case of the BUD. This approach addresses the most important shortcomings of classical asymptotic methods by explicitly consideration of the complex dependencies caused by clustering, as well as by adaptive treatment allocation.

The paper presents a new statistical model based on saddlepoint approximation to refine inference on linear rank tests based on the randomized block urn design. We support popular statistics, such as Gehan-Wilcoxon [8], Prentice–Wilcoxon [28], Mann–Whitney [32] and log-rank tests [34]. The intra-cluster correlations and adaptive allocation mechanism are modelled [22, 37] directly, we are superior to the traditional methods in small to moderate sample conditions. There are comprehensive simulations and applications of our method with real-data that confirm that it provides accurate p-values, strong confidence intervals, and better control over Type-I error rates and power [23–25].

The rest of this paper is structured in the following way: Section 2 presents the notation, the clustered data structure, and the urn randomization scheme, the weighted log-rank statistic. In section 3, the saddlepoint approximation of the test statistics is obtained. Section 4 of the paper explains how to construct confidence intervals of treatment effects. Simulation studies and real-data applications that prove the approach are provided in section 5. Lastly, Section 6 will give conclusion remarks.

2. Clustered rank-based statistic under the block urn design

In research that deals with clustered data, or multiple measurements or subunits within larger units, such as patients, family, or clinical center, it is necessary to take into account intra-cluster correlation. Failure to consider these dependencies may result in erroneous inferences, in particular Type-I errors that are overstated and misleading estimates of the variance as shown by Rosner and Grove [32]. Therefore, the rank-based test statistics are forced to be specifically adjusted to this cluster structure.

Consider a clinical trial design with K blocks, where the k-th block contains n_k clusters. Each cluster i within block k comprises g_{ki} subunits, for $i = 1, \ldots, n_k$ and $k = 1, \ldots, K$. The total number of subunits in block k is given by $G_k = \sum_{i=1}^{n_k} g_{ki}$.

The allocation of treatments within a block is controlled by a randomized urn design which is the block urn design where a block urn design is an adaptive treatment allocation that assigns subunits to treatments in a manner that maintains balance. In block k, m_{k1} subunits are assigned to treatment A, and the remaining $m_{k2} = G_k - m_{k1}$ subunits are assigned to treatment B. This is an adaptive mechanism to see that the treatment groups do not unbalance within blocks as the trial progresses sequentially.

Formally, consider a sequential trial comparing two treatments, A and B, with a target allocation vector $\mathbf{m}_k = (m_{k1}, m_{k2})$ for block k. The size of the minimal balanced set is defined as the minimal number of assignments that one needs to make the balance of the assignments complete:

$$w_k = \sum_{\varsigma=1}^{a} m_{k\varsigma}, \tag{2.1}$$

where a denotes the number of treatments (here, $a = 2$), and $m_{k\varsigma}$ is the target number of subunits for treatment ς. The total number of subunits in block k can thus be expressed as $G_k = \lambda w_k$, where λ denotes the number of minimal balanced sets that one can get in the block. This parameterization allows block sizes to be flexible and the structure of balance that is inherent to the urn design is imposed.

The explicit modeling of these cluster and block-level structures in the rank-based test statistic and its variance means that we perfectly model the dependencies both caused by the clustering of the data and the adaptive nature of the treatment allocation. The method is valid inference at small to moderate sample sizes at which the classical approximations to the asymptotic fail.

Block urn design randomization procedure: The BUD uses a dual-urn model, i.e., an active urn and an inactive urn, in creating a randomized allocation sequence that provides balance. The specific process is described in the following:

1. **Initialization:** Begin with the active urn containing G_k balls of different colors, where each color represents a treatment group $\varsigma = 1, \ldots, a$ within block k. The count of balls for color ς matches the target allocation $m_{k\varsigma}$. The inactive urn is initially empty.
2. **Random draw:** For each subject requiring assignment, randomly draw one ball from the active urn.
3. **Treatment assignment:** Assign the subject to the treatment corresponding to the color of the drawn ball.

4. **Ball transfer:** Following assignment, transfer the drawn ball from the active urn to the inactive urn.

5. **Balanced set check:** After each transfer, verify whether the inactive urn contains a complete minimal balanced set—specifically, exactly $m_{k\varsigma}$ balls for each treatment color ς.

6. **Urn reset:** If a minimal balanced set is detected in the inactive urn, immediately return these $w_k = \sum_{\varsigma=1}^{a} m_{k\varsigma}$ balls to the active urn. Any residual balls remain in the inactive urn.

7. **Repeat:** Iterate this process—drawing, assigning, transferring, and checking—until all subjects in the block have been randomized.

This process ensures that tasks are not skewed in blocks but the allocation is stochastic. The mechanism of the minimal balanced set, the interaction between the active and inactive urns, helps to avoid the imbalance build-up and maintain the desired ratios of allocation.

Rosenberger and Lachin [30] described the permuted block design as a repeated randomization, in other words an urn model with no replacement in a block. The permuted block design has balls drawn without replacement until the block is exhausted, and the balls are only returned (in principle) at the end of the block. The BUD is essentially different in its mechanism of return: It puts balls back into the active urn immediately a minimal balanced set has been built up in the inactive urn, and not when the block is complete. This minor difference changes the nature of changes in the allocation probabilities with time. It is interesting to note that in the special case when there is only one minimal balanced set in a block ($\lambda = 1$), the BUD becomes degenerated to the usual permuted block design. Similar to the permuted block design, the BUD is flexible to multi-treatment trials and can also be used with the balanced and unbalanced ratios of allocation. To treat the urn model that is underlying this process in detail, see Zhao and Weng [40].

Clustered rank-based test statistic under the block urn design: In the block urn design, the clustered rank-based test statistic is clearly defined as the weighted sum of scores of treated subjects:

$$T = \sum_{k=1}^{K} b_k \sum_{i=1}^{n_k} \sum_{j=1}^{g_{ki}} u_{kij} X_{kij}, \tag{2.2}$$

where K denotes the total number of blocks, and b_k is the weight assigned to block k ($k = 1, \ldots, K$). Within block k, n_k represents the number of clusters, and g_{ki} is the number of subunits in cluster i. The term u_{kij} is the centered score function associated with the outcome of the j-th subunit in the i-th cluster of block k. The random variable X_{kij} serves as the treatment assignment indicator:

$$X_{kij} = \begin{cases} 1, & \text{if the subunit is assigned to treatment A,} \\ 0, & \text{if the subunit is assigned to treatment B.} \end{cases}$$

Under the null hypothesis of no treatment effect, the score functions u_{kij} are regarded as fixed constants. Consequently, the stochastic nature of the test statistic T derives entirely from the random treatment assignments X_{kij}. Given a balanced allocation within each block, the expected value of the treatment indicator is $\mathbb{E}(X_{kij}) = m_{k1}/G_k$, where m_{k1} is the number of subunits assigned to treatment A in block k, and $G_k = \sum_{i=1}^{n_k} g_{ki}$ is the total block size. Since the scores are centered such that $\sum_{k=1}^{K} \sum_{i=1}^{n_k} \sum_{j=1}^{g_{ki}} u_{kij} = 0$, the expected value of the test statistic under the null is zero: $\mathbb{E}(T) = 0$.

Variance of the test statistic T: As the number of blocks $K \rightarrow \infty$, the statistic T follows an asymptotic normal distribution with mean zero and variance:

$$\mathrm{Var}(T) = \sum_{k=1}^{K} \frac{m_{k1}(G_k - m_{k1})}{G_k(G_k - 1)} b_k^2 \sum_{i=1}^{n_k} \sum_{j=1}^{g_{ki}} u_{kij}^2.$$

This term presupposes the lack of correlation between blocks and fixed marginal totals on each block [39]. It is suitable to explain the stratified nature of the data and the limitations of the block urn randomization. As stressed by Rosner and Grove [32], failure to take into account such clustering effects can generally result in deflated estimates of variance, inflated Type-I error rates and faulty statistical inference. Although the prediction of the extended form of the variance estimator uses four different parameters of correlation to reflect all the complicated inter-cluster dependencies, the simplified version above is the standard estimator of randomized block designs with equal allocation as used to derive the asymptotic benchmark in this paper.

Standardized test statistic and asymptotic distribution: Hypothesis testing proceeds by standardizing the test statistic:

$$Z = \frac{T}{\sqrt{\mathrm{Var}(T)}}.$$

Under the null hypothesis, the standardized statistic is, in distribution, equal to the standard normal as K increases: $Z \xrightarrow{d} \mathcal{N}(0,1)$. This is an asymptotic value that makes it easy to compute p-values and confidence intervals. In practice, the components of variance are always computed based on the observed data making the test resistant to unbalanced cluster sizes or blocking sizes-which are typical of real-life clinical data. As an example, location shift alternatives can be made efficient by determining block weights as $b_k = (G_k + 1)^{-1}$ when using Wilcoxon scores, as observed by Van der Waerden [35]. This structure provides valid inference and better statistical characteristics than naive techniques that do not incorporate the randomization characteristics of the design or the fact that data are clustered boundaries.

Clustered rank-based tests for survival data: When the survival data is clustered and there is no censoring then the rank statistics used to define the statistics in the form of a class are as given in Eq (2.2) also includes the clustered Mann-Whitney U-test. After the formulation by Rosner and Grove [32], the score function u_{kij} for the j-th subunit in cluster i of block k assigned to treatment A is constructed as:

$$u_{kij} = \sum_{k',\ell,m} \left[I(X'_{kij} < Y'_{k'\ell m}) + \frac{1}{2} I(X'_{kij} = Y'_{k'\ell m}) \right], \tag{2.3}$$

where X'_{kij} represents the outcome of the j-th subunit in cluster i of block k (treatment A), and $Y'_{k'\ell m}$ represents the outcome of the m-th subunit in cluster ℓ of block k' (treatment B). The indicator function $I(\cdot)$ equals 1 if the condition holds and is 0 otherwise. Essentially, this score quantifies the number of observations in the control group (treatment B) that exceed the value of X'_{kij}, with ties contributing a weight of 0.5. Alternatively, this score can be equivalently expressed using the centered rank of X'_{kij} within the pooled sample of N total observations:

$$u_{kij} = R_{kij} - \frac{N+1}{2}, \tag{2.4}$$

where R_{kij} is the rank of X'_{kij} among all observations across both treatment groups. Centering the ranks by subtracting $(N+1)/2$ defines the sum of the scores to be equal to zero, which is important to the asymptotic behaviour and symmetry of the test statistic.

Extension to survival data with censoring: Where there exists censoring, the score definition should take into consideration incomplete survival times. Following Jeong and Jung [12], the score function for the j-th subunit in cluster i assigned to treatment group k is adapted as:

$$u_{ikj} = \delta_{ikj}\, G(X^*_{ikj}) \frac{Y_{3-k}(X^*_{ikj})}{Y(X^*_{ikj})}, \tag{2.5}$$

where X^*_{ikj} is the observed survival time (censored or event), and δ_{ikj} is the event indicator (1 if the event is observed, 0 if censored). The term $Y(t)$ denotes the total number of individuals at risk at time t, while $Y_{3-k}(t)$ denotes the number at risk specifically in the *opposing* treatment group. The weight function $G(t)$ determines the specific type of rank test:

- **Log-rank test:** $G(t) = \frac{1}{n}\frac{Y_1(t)Y_2(t)}{Y(t)}$, emphasizing differences across the survival curve equally.
- **Gehan–Wilcoxon test:** $G(t) = \frac{1}{n^2}Y_1(t)Y_2(t)$, which places greater weight on early events.
- **Prentice–Wilcoxon test:** $G(t) = \frac{1}{n}\hat{S}^-(t)\frac{Y_1(t)Y_2(t)}{Y(t)}$, using the left-continuous pooled Kaplan–Meier estimator $\hat{S}^-(t)$ [13] to modify the weighting scheme.

Null hypotheses for clustered rank tests: In the case of the Mann-Whitney test, the null hypothesis is that the two treatment groups have the same marginal distribution of the outcome of interest:

$$H_0 : F_1(x) = F_2(x) \quad \text{for all } x,$$

where $F_1(x)$ and $F_2(x)$ are the respective cumulative distribution functions (CDFs). Equivalently, this can be stated probabilistically as $H_0 : \Pr(X < Y) = 0.5$, where X and Y are independent random draws from groups 1 and 2.

Under survival analysis (as in Jeong and Jung [12]), the null hypothesis is that marginal survival functions are equal as a time-dependent:

$$H_0 : S_1(t) = S_2(t) \quad \text{for all } t > 0,$$

where $S_k(t)$ indicates the marginal survival function of treatment group k. These score functions give a flexible method of testing the treatment effect in complex and correlated data environments, which supports both uncensored and censored outcomes and strictly takes into account the cluster structure.

Remark 2.1. *The score functions defined in Eq (2.5) targets standard right-censored data, which in the case of competing risks is equivalent to testing differences in cause-specific hazards. These scores need to be redefined to test hypotheses about cumulative incidence functions (e.g., by the weighting method suggested by Gray [9]). More importantly, though, the saddlepoint framework in Section 3 is independent of the choice of scores. Since the approximation is based on the conditional randomization distribution of assignments, conditioned on fixed scores, it is just necessary to replace the competing-risk scores with the statistic T and the resulting saddlepoint approximation continues to be valid and computationally exact in the new context.*

3. Saddlepoint approximations in the context of randomized block urn design

Saddlepoint approximations are a solid model of estimating the distribution of test statistics with great accuracy. This approach is especially useful in complicated experimental designs where classical asymptotic theory is frequently ineffective, e.g., where the sample sizes are small, censoring is high or the dependencies within clusters are complicated. Saddlepoint methods can effectively solve these inferential problems by offering much more accurate estimates of tail probabilities and distributional approximations of even the extreme tails [19, 22, 23].

These approximation techniques are not only useful but necessary in the particular situation of the BUD. They play a vital role in being highly rigorous in dealing with the statistical complexities that are posed by adaptive treatment allocation and associated dependencies in clustered data. It is a systematic extension of saddlepoint approximations to BUD, which is far broader in scope than earlier approaches to generalized randomized block designs (e.g., [1, 20, 21]). Our methodology is fully supportive of the special stochastic form of urn-based randomization and nested clustering.

The BUD uses a dynamic urn model in order to modify the treatment assignment probabilities sequentially. Although such a mechanism guarantees the best balance, it also creates certain stochastic dependencies both between and within blocks. Contrary to traditional block designs (fixed size) where the assignments have been determined in advance, the BUD will update the composition of the urn at each randomization of a subunit. This non-independent dynamic distribution requires an accurate approximation of the test statistic finite-sample distribution. Standard normal approximations are not always reliable when dealing with small-to-moderate sample sizes as used in clinical trials, especially the tail areas where the calculation of the p-value is required. Saddlepoint algorithms do not have these shortcomings, as the cumulant generating function (CGF) appropriately represents the specific probabilistic structure of the adaptive design.

To formalize this framework, consider the generalized linear rank test statistic T, constructed as a weighted sum of treatment assignments and rank scores. Let K denote the number of blocks, n_k the number of clusters in block k, and g_{ki} the number of subunits within cluster i of block k. Let X_{kij} represent the binary treatment assignment indicator (1 for treatment A, 0 for B) for the j-th subunit within the i-th cluster of the k-th block. Under the BUD, the joint randomization distribution of the assignment vector $\mathbf{X} = (X_{111}, \ldots, X_{Kn_Kg_{Ki}})$ is defined conditionally on the realized block totals:

$$\mathbf{X} \overset{D}{\sim} \mathbf{x} \left| \bigcap_{k=1}^{K} \left\{ \sum_{i=1}^{n_k} \sum_{j=1}^{g_{ki}} x_{kij} = m_{k1} \right\} \right. .$$

Here, $\overset{D}{\sim}$ denotes "distributed as", and m_{k1} represents the fixed number of subjects assigned to treatment group 1 in block k over the entire trial. Crucially, the individual assignment probabilities $p_{kj\varsigma}$ the parameters to govern these indicators are not fixed values of Bernoulli but they are produced dynamically through the mechanism of the urn.

We condition on the vector of realized allocation counts $\mathbf{m} = (m_{11}, \ldots, m_{K1})$ to adhere to the principle of conditional inference. Since the block urn design generates random group sizes, unconditional inference would consist of averaging over sample size imbalances which were not observed and which might over inflate the variance. By fixing $\mathbf{m}$, we limit the set of reference to the permutations consistent with the observed trial properties, and make the allocation counts auxiliary statistics. However, the specific probabilistic dynamics of the urn are preserved within this conditional framework through the individual $p_{kj\varsigma}$ terms in the CGF.

The linear rank test statistic T is defined as:

$$T = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \sum_{j=1}^{g_{ki}} U_{kij} X_{kij},$$

where $U_{kij} = b_k u_{kij}$ are the rank scores, which are predetermined. The aim of the saddlepoint approximation here is the randomization distribution of T over the null hypothesis with the counts being fixed m_{k1}:

$$P(T \le t \mid \mathbf{m}) = P\left(\sum_{k=1}^{K} \sum_{i=1}^{n_k} \sum_{j=1}^{g_{ki}} U_{kij} X_{kij} \le t \;\middle|\; \bigcap_{k=1}^{K} \left\{ \sum_{i=1}^{n_k} \sum_{j=1}^{g_{ki}} X_{kij} = m_{k1} \right\} \right).$$

In the BUD, the probability $p_{kj\varsigma}$ that the next subject is assigned to treatment ς depends directly on the current composition of the active urn. For a balanced two-treatment case ($\varsigma = 1, 2$), the conditional probability for subject j (omitting block indices k, i for clarity) is typically formulated as:

$$p_{kj1} = \frac{\lambda + \min(\nu_{j-1,1}, \nu_{j-1,2}) - \nu_{j-1,1}}{2\lambda + 2\min(\nu_{j-1,1}, \nu_{j-1,2}) - (j-1)}, \quad \text{and} \quad p_{kj2} = 1 - p_{kj1},$$

where:

- λ: A design parameter (e.g., $\lambda = 1$) representing the initial urn composition. Higher λ values reduce the adaptive strength, approaching complete randomization.
- $\nu_{j-1,\varsigma}$: The cumulative count of subjects assigned to treatment ς prior to the current subject j.

These dynamic probabilities induce complex, non-standard dependencies that saddlepoint methods are uniquely equipped to handle.

The mid-p-value: Refining significance for discrete data: The critical refinement of hypothesis testing using discrete test statistics is the mid-p-value [14]. In comparison with the traditional p-value which is strictly conservative in discrete distributions, the mid-p-value incorporates one half of the probability of the outcome observed. The resulting correction has a more balanced average value of a significance measure that is nearer to the nominal level, enhances Type-I error control, but is not anti-conservative [6]. Mathematically, given an observed test statistic value t the mid-p-value is defined as:

$$\text{mid-}p = P(T < t) + \frac{1}{2} P(T = t).$$

This expression is necessary in the event that the outcome data are tied by ties which result in tied rank scores (u_{kij}) thus making the distribution of T to be lattice-like instead of continuous. The mid-p-value corrects the discrete nature of ties and finite sample space by adding half the probability mass at the observed value. Although this value can be approximated by simulation-based methods (i.e., with replications of say, 10^6), it is computationally expensive. Saddlepoint approximations provide a very accurate alternative of analytical evaluation.

Saddlepoint approximation of the mid-p-value: We utilize the double saddlepoint formula derived by Skovgaard [33] and refined by Booth and Butler [2] for conditional distributions. This is a formula that has been specially designed to deal with discrete variables and estimates the mid-p-value directly:

$$\text{mid-}p \approx H(\widehat{\vartheta}) + h(\widehat{\vartheta})\left(\frac{1}{\widehat{\vartheta}} - \frac{1}{\widehat{\rho}}\right), \tag{3.1}$$

where H and h are the standard normal CDF and probability density function (PDF), respectively. The deviates $\widehat{\vartheta}$ and $\widehat{\rho}$ are defined as:

$$\widehat{\vartheta} = \text{sgn}(\widehat{t})\left[2\left\{\left(K(0,\widehat{S}_0) - M^T\widehat{S}_0\right) - \left(K(\widehat{t},\widehat{S}) - M^T\widehat{S} - \zeta_0\widehat{t}\right)\right\}\right]^{1/2},$$

$$\widehat{\rho} = \widehat{t}\left[\frac{|K''(\widehat{t},\widehat{S})|}{|K''(0,\widehat{S}_0)|}\right]^{1/2}.$$

The components of this approximation are:

- $M = (m_{11}, \ldots, m_{K1})^T$: The vector of observed assignments to treatment group 1 in each block.
- $\widehat{S}$ and $\widehat{S}_0$: Vectors of nuisance parameters (Lagrange multipliers) solved from the saddlepoint equations at $\widehat{t}$ and $t = 0$, respectively.
- $K(t, S)$: The joint CGF of the test statistic T and the block sums. For the BUD with binary outcomes, this is:

  $$K(t, S) = \sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{j=1}^{g_{ki}} \ln\left[(1 - p_{kj1}) + p_{kj1}\exp(s_k + tU_{kij})\right].$$

  Here, p_{kj1} captures the adaptive urn probability.

  Note: According to this formulation, the design is not ignored. The CGF combines conditional Bernoulli trials with the urn history as opposed to assuming that the assignment is independent and identical distributed (i.i.d.). We compute the adaptive nature of the design using the saddlepoint equations to condition on observations of the design, which are m_{k1}, and obey the constraints of the achieved sample size. The u_{kij} scores are constant numbers (using mid-ranks to treat ties), so that the approximation will automatically reflect the discrete nature of the data.
- $K''(t, S)$: The Hessian matrix of the CGF, used to adjust for curvature.

- The saddlepoint $\widehat{t}$ and vector $\widehat{S}$ are the unique solutions to:

$$\frac{\partial K}{\partial t}(\widehat{t}, \widehat{S}) = \zeta_0, \quad \text{and} \quad \frac{\partial K}{\partial s_k}(\widehat{t}, \widehat{S}) = m_k, \quad k = 1, \ldots, K.$$

Similarly, $\widehat{S}_0$ satisfies these equations when $t = 0$.

The solution of these equations gives the saddlepoint values that identify the most likely state of the data that is consistent with the observed tail event, which gives very accurate p-values.

Remark 3.1. *Although the usual p-value is conventional, in the discrete case the saddlepoint approximation is theoretically associated with the mid-p-value. The formula is an implicit continuity correction which smooths the lattice distribution of the test statistic. Therefore, the approximation error is smaller when one wants to achieve the mid-p-value (which is $(O(n^{-1})))$ instead of the actual tail probability $(P(T \geq t))$ and the error is dominated by the discontinuity of the data. Therefore, the mid-p-value is not only a preference of a metric of significance but also a prerequisite to the analytical accuracy of the saddlepoint method.*

Remark 3.2. *The formulation of the CGF $K(t, S)$ relies exclusively on the randomization probabilities p_{kj1} defined by the block urn design. It does not make any assumptions regarding the intra-cluster correlation structure of the outcomes (e.g., gamma frailty). Since the saddlepoint approximation is conditional on the observed data, any complicated or even nested dependencies in the outcomes will be directly reflected by the fixed score values u_{kij} by definition. Therefore, the CGF can be used on general dependence structures with no changes provided that the randomization protocol is kept fixed.*

4. Confidence intervals for treatment effect: Inverting the rank test

In the analysis of survival data, the treatment effect is often conceptualized within the accelerated failure time (AFT) framework as a location shift parameter β on the log-time scale [38]. To construct confidence intervals (CIs) for β using linear rank tests, we employ the principle of test inversion [14]. This procedure identifies the set of all hypothesized treatment effects β_0 for which the null hypothesis $H_0 : \beta = \beta_0$ is not rejected at a specified significance level α.

The inversion process for a specific candidate β_0 involves the following steps, see [23–25]:

1. **Pseudo-observation creation:** Let $y_{kij} = \log(t_{kij})$ denote the observed log-transformed survival or censoring time for the j-th subunit in cluster i of block k. For subjects in the treatment group ($X_{kij} = 1$), we adjust the observed times by subtracting the hypothesized effect: $y^*_{kij} = y_{kij} - \beta_0$. For subjects in the control group ($X_{kij} = 0$), the observed times remain unchanged ($y^*_{kij} = y_{kij}$). This transformation aligns the distributions of the two groups under the null hypothesis that β_0 is the true effect.

2. **Rank-based score calculation:** Rank scores, denoted as $u_{kij}(\beta_0)$, are computed based on the joint ranking of these adjusted observations y^*_{kij}. The calculation must rigorously account for censoring, utilizing standard survival score functions such as the log-rank, Gehan–Wilcoxon, or Prentice–Wilcoxon weights [38].

More importantly, score definitions in clustered designs such as the BUD should also consider intra-cluster correlation to eliminate deflation of variance and invalid inference. The best way to do it is to get marginal model scores by explicitly modeling within-cluster dependence, including marginal Cox models with frailty terms (random effects) [15,22]. Although the overall theory of frailty-based scores is quite accepted [19–21], calculating $u_{kij}(\beta_0)$ for rank tests involves applying these principles to the adjusted data y^*_{kij}. This guarantees that the scores capture the treatment effect that is being hypothesized and also the cluster structure.

With the β_0-dependent scores established, the test statistic $T(\beta_0)$ is computed. The corresponding p-value is then evaluated using the mid-p formulation derived in Section 3:

$$\widehat{P}(\beta_0) = P\left(T(\beta_0) < t_{obs}(\beta_0)\right) + \frac{1}{2}P\left(T(\beta_0) = t_{obs}(\beta_0)\right),$$

where $t_{obs}(\beta_0)$ is the realized value of the test statistic for the adjusted data. The $100(1-\alpha)\%$ confidence interval for β comprises all values of β_0 such that $\widehat{P}(\beta_0) \in [\alpha/2, 1-\alpha/2]$. Numerically, this interval is determined by searching over a grid of plausible β_0 values (e.g., with step size 0.001). The iterative procedure is summarized as follows:

1. Adjust log-times to generate y^*_{kij} for the current β_0.
2. Re-rank the pooled sample, accounting for censoring.
3. Recalculate scores $u_{kij}(\beta_0)$ to capture the updated ranks and clustering.
4. Compute the test statistic $T(\beta_0)$.
5. Calculate the mid-p-value using the saddlepoint approximation.

This is done by monotonicity of the test statistic as the relative ranking of the treatment and control observations changes with the change in the value of β_0, which enables the exact determination of the limits of the confidence interval.

Harnessing saddlepoint accuracy for confidence intervals: We employ the saddlepoint approximation in order to calculate the necessary mid-p-values with efficiency. The reason why this method is desirable compared to normal approximations of the standard form is due to the fact that it gives better accuracy on the discrete skewed distributions that are characteristic of the rank tests [4]. When the sample sizes are small, and dependencies are complicated, as is the case in the randomized block urn design, the nominal coverage is frequently not maintained by the standard large-sample approximations, especially in the tails.

It has been shown in previous literature e.g. the classic leukemia remission data analyzed by [1, 8], that saddlepoint-based confidence intervals match the exact permutation intervals, and the normal approximations often have inflated widths or coverage errors. Here we are hoping to have the same benefits. The saddlepoint method provides rigorous Type-I error control at each grid search point by correctly characterizing the exact conditional randomization distribution, which is the urn mechanism dependencies and clustering dependencies. Simulation studies have been given in the following section to confirm these properties, including empirical coverage probabilities and interval widths on a variety of clinical trial situations.

5. Simulation study and real-data applications

5.1. Simulation study

In order to conduct a systematic analysis of the effectiveness of the proposed saddlepoint approximation of clustered rank tests using the block urn design, we performed a large scale simulation study. It is mainly aimed at measuring the accuracy and strength of the method compared to traditional asymptotic methods, especially when the sample size is small to moderate size and mixed adaptive dependencies [3].

For each simulation replicate, the rank test statistic T (Eq (2.2)) was calculated, and inference was performed using three distinct methodologies:

1. Proposed saddlepoint approximation (SP p-value): This is the main contribution we make. The saddlepoint p-value as calculated by reverse conditional CGF of the conditioned test statistic as derived in Section 3 looks as follows. This approach implicitly takes into consideration skewness and kurtosis brought about by the adaptive urn mechanism and intra-cluster correlation by using the full functional form of the CGF. It produces the most precise tail probabilities of discrete non-normal distributions, and produces values that are close approximations to the actual conditional mid-p-value.

2. Asymptotic normal approximation (AN p-value): In this standard method, the calculation of the p-values is done by standardizing T with the robust variance estimator obtained in Section 2 and comparing the value with the standard normal distribution. Although it is computationally easy, its validity is based on large-sample central limit theorems. Our simulations should prove that this dependency frequently results in inflated Type-I error rates and in poor coverage of the confidence interval when the sample size is not very large and when the dependencies caused by the urn are very strong.

3. Simulated mid-p-value (Benchmark): In order to define a ground truth on accuracy, we computed an empirical mid-p-value with a large amount of Monte Carlo simulation. In each replicate, the number of permutations generated conditional on the observed counts of blocks were 10^6 to give a loose approximation of the true randomization distribution with a standard error that is negligible. The benchmark p-value is defined as $\widehat{P}_{\text{sim}} = P(T < t_{obs}) + 0.5P(T = t_{obs})$ [22]. We expect the proposed SP p-value to align almost perfectly with this benchmark, thereby validating its theoretical exactness.

5.1.1. Simulation design and data generation

The simulation study aims to strictly reproduce the complexities that are presented in clinical trials especially those that involve clustered data and adaptive randomization. The scenarios are each assessed with a number of independent replication of $N_{sim} = 20,000$, which is sufficient to determine exactly the Type-I error rates, statistical power, and coverage of the confidence interval (margin of error of about $\approx \pm 0.005$). For each replicate, data generation proceeds in three stages:

1. Clustered outcome simulation: We generate outcomes for individual subunits with explicit intra-cluster correlation (ρ).

- For time-to-event data, we employ shared frailty models [10, 18]. Such models add a latent cluster-specific random effect (frailty) that multiplicatively acts on the baseline hazard and brings about dependence among subunits of the same cluster. To test robustness we change the baseline distributions (exponential when the hazard does not change with time and Weibull when it does).
- In continuous non-survival data (i.e. when we are using Mann-Whitney U-test), we employ linear mixed models: $Y_{kij} = \mu + U_i + \epsilon_{kij}$, where $U_i \sim N(0, \sigma_U^2)$ is the random cluster intercept and $\epsilon_{kij} \sim N(0, \sigma_\epsilon^2)$ is the residual error. The intra-cluster correlation is defined as $\rho = \sigma_U^2/(\sigma_U^2 + \sigma_\epsilon^2)$.

2. Randomized allocation: Assignments (A or B) of treatment are generated in a sequencing way by use of the BUD presented in Section 2. This step brings about the history-dependent allocation probabilities which we are going to be investigating.
3. Censoring mechanism: In the case of survival endpoints, independent right-censoring times are based on uniform distributions with constants adjusted to produce desired censoring rates which can simulate administrative censoring or the loss to follow-up.

Generating clustered survival data via shared frailty models: The algorithm that is used in simulating clustered survival data is as shown below 1. The most important parameters that delineate the data structure and correlation are:

1. **Trial design structure:**
   - K: Number of blocks.
   - n_k: Number of clusters per block (fixed or random, e.g., $U[2, 5]$).
   - g_{ki}: Cluster size (number of subunits, fixed or random, e.g., $U[3, 7]$).
2. **Frailty distribution** (Z_i)**:** We assume gamma-distributed frailties, $Z_i \sim \text{Gamma}(\kappa, \kappa)$, with mean 1 and variance $1/\kappa$. The dependence parameter κ relates to Kendall's τ and the intra-cluster correlation. We consider three levels of dependence:
   - Low: $\rho \approx 0.2$ ($\kappa = 4$).
   - Moderate: $\rho \approx 0.5$ ($\kappa = 1$).
   - High: $\rho \approx 0.7$ ($\kappa \approx 0.43$).
3. **Baseline hazard $\lambda_0(t)$:**
   - Exponential: $\lambda_0(t) = \lambda_0$ (constant risk).
   - Weibull: $\lambda_0(t) = (\eta/\nu)(t/\nu)^{\eta-1}$ (increasing/decreasing risk).
4. **Treatment effect** (β)**:** Modeled as a log-hazard ratio. For power simulations, we set specific hazard ratios (HRs); for example, $\text{HR} = \exp(-0.5) \approx 0.606$ corresponds to a moderate treatment benefit.

Algorithm 1 Simulation algorithm for clustered rank tests

```
Input: Design parameters $(K, n_k, g_{ki})$, frailty parameter $(\kappa)$, baseline hazard parameters ($\lambda_0$ or $\eta, \nu$), log-hazard ratio $(\beta)$, censoring distribution.
Generate frailties: Draw $Z_i \sim \text{Gamma}(\kappa, \kappa)$ for all clusters $i = 1, \dots, n_k$ in blocks $k = 1, \dots, K$.
Assign treatments: Generate treatment indicators $x_{kij}$ and block counts $m_{k1}$ using the BUD procedure, see Section 2.
Simulate survival times $(T_{kij})$:
Calculate subject-specific hazard: $\lambda_{kij}(t) = Z_i \cdot \lambda_0(t) \cdot \exp(\beta \cdot x_{kij})$.
Draw $U \sim \text{Uniform}(0, 1)$.
if Exponential Baseline then
    $T_{kij} = -\ln(U)/\lambda_{kij}$.
else if Weibull Baseline then
    $T_{kij} = \nu \left( \frac{-\ln(U)}{Z_i \exp(\beta x_{kij})} \right)^{1/\eta}$.
end if
Simulate censoring $(C_{kij})$: Draw $C_{kij}$ from specified distribution to match target censoring rate.
Determine observed data:
Time: $T^*_{kij} = \min(T_{kij}, C_{kij})$.
Status: $\delta_{kij} = I(T_{kij} \le C_{kij})$.
Compute statistic: Calculate observed rank statistic $T_{obs}$ using $(T^*_{kij}, \delta_{kij})$ and block weights $b_k$.
Perform inference: Compute SP $p$-value, AN $p$-value, and benchmark mid-$p$-value.
```

In non-survival cases, when we are using the Mann-Whitney U-test, we use the linear mixed model, which is given above. These normal variates are exponentiated to obtain lognormal data. Such a detailed configuration will enable the consideration of the method performance regarding a wide range of distributional shapes and correlation configurations.

The illustrative simulation scenarios have particular configurations, which are described in Table 1. We evaluate the effectiveness of both the methods according to the following measures:

- **Type-I error rate:** In the null hypothesis ($\Delta = 0$), we find the percentage of replicates in which the p-value is not greater than ≤ 0.05. Strong techniques are expected to produce rates that are near the nominal level (Table 2).

- **Accuracy mean absolute error:** To measure the accuracy of the approximation, we use the mean absolute error (MAE) between the proposed SP p-value and the benchmark simulated mid-p-value (Table 3).

- **Coverage probability:** In the case of 95% confidence intervals, we report the fraction of intervals which include the true parameter, β. Preferably, this is supposed to be near 0.95.

- **Interval width:** The mean width of the confidence intervals are reported to determine precision; smaller width (with the right coverage) means that it is more efficient (Table 4).

Visual comparisons of these metrics are provided in Figures 1–3.

Table 1. Key simulation parameters for illustrative scenarios.

Parameter	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6
Urn design parameters						
γ (Balance factor)	0.5	1.0	0.5	0.5	1.0	0.5
α (Adaptive factor)	1.0	0.5	1.0	2.0	1.0	1.0
Clustered data structure						
Blocks (K)	10	15	20	10	15	20
Clusters/Block (n_k)	Ran. (2–5)	Ran. (3–6)	Fix. (3)	Ran. (2–4)	Fix. (4)	Ran. (3–7)
Subunits/Cluster (g_{ki})	Ran. (3–7)	Fix. (5)	Ran. (4–8)	Fix. (5)	Fix. (6)	Fix. (4)
ICC (ρ)	0.2	0.4	0.6	0.3	0.5	0.7
Total N (approx.)	150–350	225–450	240–480	100–200	360–540	240–560
Treatment effect (Δ) for outcomes						
Null hypothesis (Type-I)	0	0	0	–	–	–
Alternative (Power)	–	–	–	0.5 (LR)	1.0 (MWU)	0.75 (Gehan-W.)
Data distribution						
Type (Survival data)	Exp.	Exp.	Weibull	Exp.	Lognormal	Weibull
Distribution parameters	0.05	0.03	(sh 1.5, sc 100)	0.04	(μ=5, σ=1)	(sh 1.2, sc 80)
Censoring rate (%)	25	40	30	20	35	45
Tests evaluated	LR, MWU	LR, Prem.	LR, Gehan	LR	MWU	Gehan-W.
Replications (N_{sim})	20,000 for each scenario					

Cl.: clusters; **Subunits/Cluster:** number of subunits per cluster; **Fix.:** fixed; **Ran.:** random; **ICC:** intra-cluster correlation; **LR:** log-rank test; **MWU:** Mann–Whitney U-test; **Prem.:** Prentice–Wilcoxon test; **Gehan-W.:** Gehan–Wilcoxon test; **Exp.:** exponential distribution; **Lognormal:** lognormal distribution; **sh/sc:** Weibull shape/scale; N_{sim}: number of simulation replications.

Table 2. Type-I error rates (nominal α = 0.05).

Scenario	Test Stat.	Proposed SP	Asymptotic Normal	Simulated Mid-p
1 (ρ = 0.2, K = 10)	Log-rank	0.052	0.118	0.051
	Mann-Whitney U	0.049	0.105	0.050
2 (ρ = 0.4, K = 15)	Log-rank	0.051	0.132	0.051
	Prentice-Wilcoxon	0.053	0.125	0.052
3 (ρ = 0.6, K = 20)	Log-rank	0.048	0.141	0.049
	Gehan-Wilcoxon	0.050	0.135	0.050

Table 3. Statistical power (nominal α = 0.05).

Scenario	Test Stat.	True Effect (Δ)	Proposed SP	Normal	Simulated Mid-p
4 (ρ = 0.3, K = 10)	Log-rank	0.50	0.765	0.652	0.763
5 (ρ = 0.5, K = 15)	Mann-Whitney U	1.00	0.821	0.701	0.820
6 (ρ = 0.7, K = 20)	Gehan-Wilcoxon	0.75	0.798	0.685	0.720

Table 4. Confidence interval performance (nominal 95% CI).

Scenario	Test Stat.	True Effect (Δ)	Empirical Coverage			Average CI Width		
			SP	Normal	Mid-p	SP	Normal	Mid-p
4	Log-rank	0.50	0.948	0.902	0.947	0.85	1.02	0.83
5	Mann-Whitney U	1.00	0.951	0.895	0.953	1.35	1.68	1.36
6	Gehan-Wilcoxon	0.75	0.949	0.908	0.943	1.12	1.37	1.12

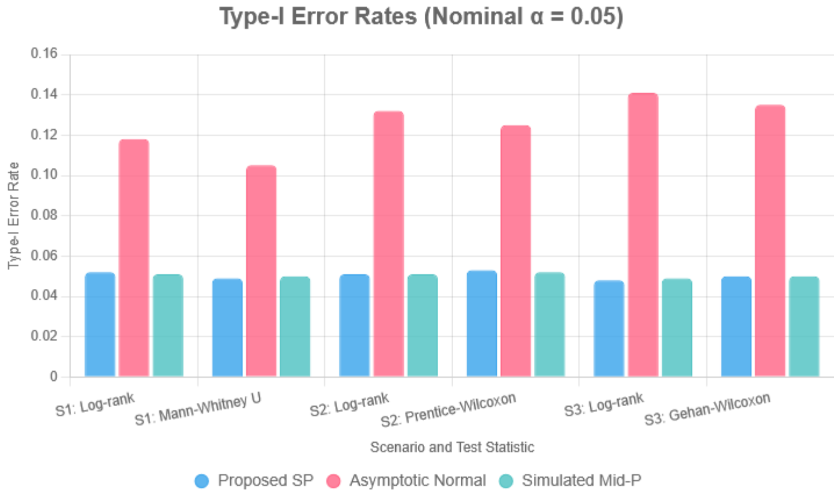


Figure 1. Type-I error rates across scenarios (nominal $\alpha = 0.05$). The bar chart is a comparison of empirical Type-I error rates when the proposed saddlepoint (blue), asymptotic normal (red) and simulated mid-p (turquoise) are used in Scenario (1,2,3) and test statistics when testing a null hypothesis. The suggested saddlepoint algorithm always keeps Type-I error rates at or close to the nominal 0.05 threshold, which is similar to the simulated mid-p benchmark. Contrastingly, Type-I error rates are highly inflated in the asymptotic normal approach (they can be quite high, e.g. above 0.10), which means that the probability of false positives when assumptions are not met is unacceptable.

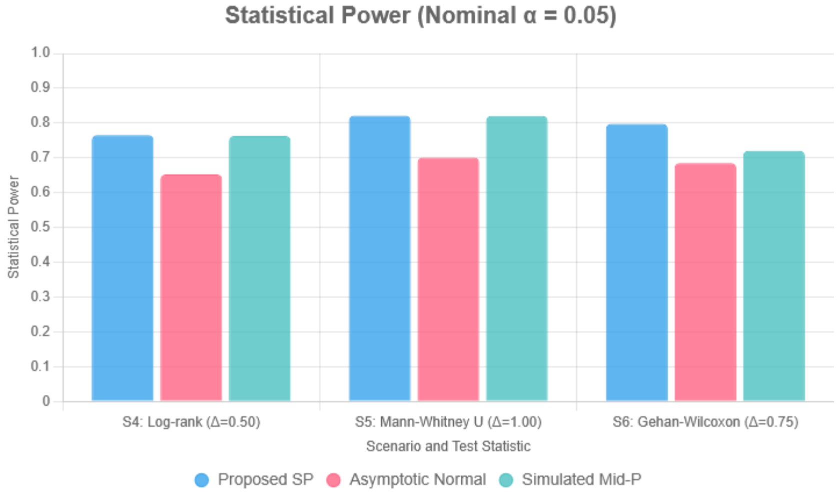
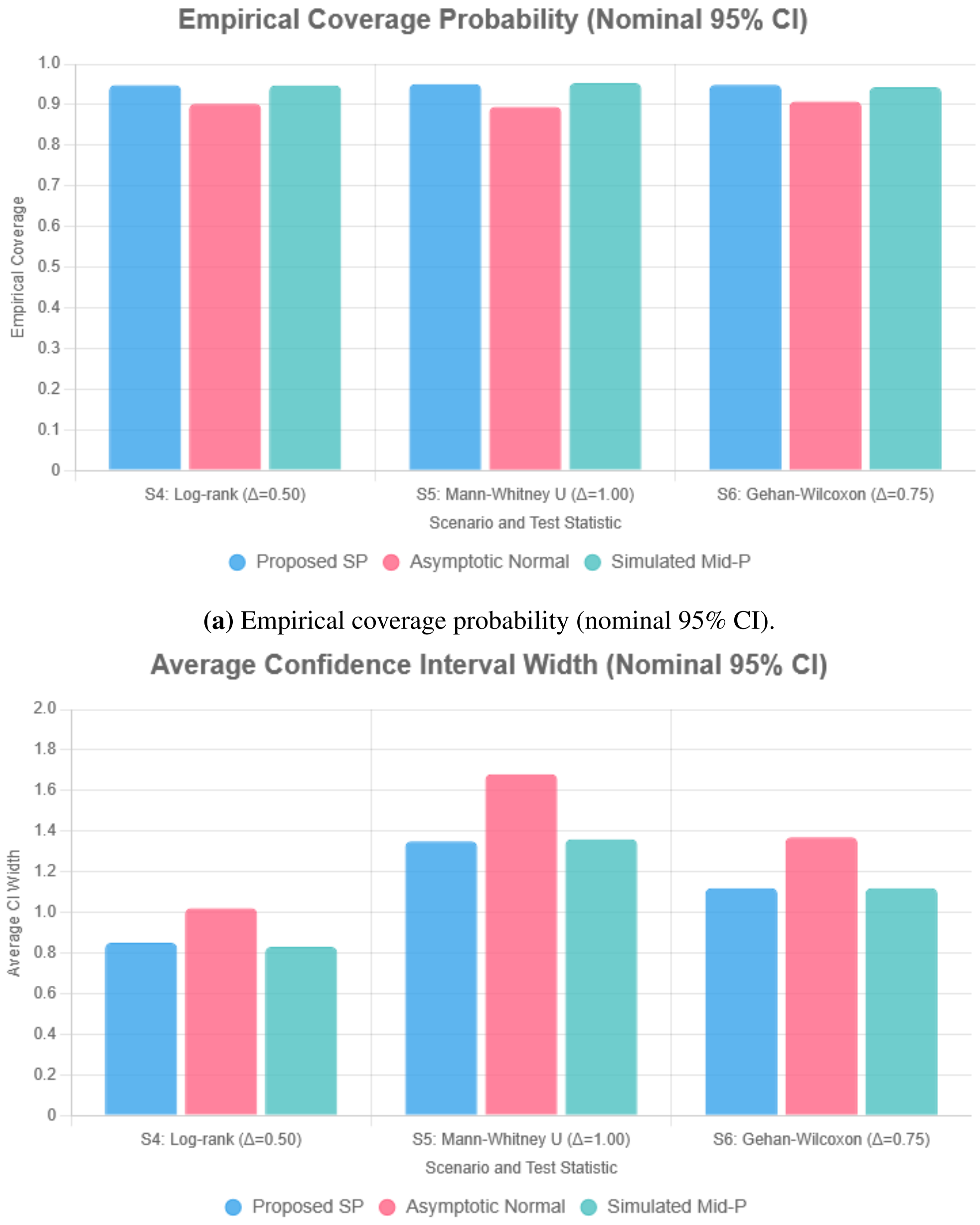


Figure 2. Statistical power across scenarios (nominal $\alpha = 0.05$). This bar graph shows the empirical statistical power of the proposed saddlepoint (blue), asymptotic normal (red), and simulated mid-p (turquoise) methods with different alternative hypotheses (Scenarios 4,5,6). We have shown that the proposed saddlepoint method is always more powerful than the asymptotic normal method, meaning that it is more capable of identifying true treatment effects in a variety of different situations and test statistics. The strength of the simulated mid-p benchmark is also close to or even greater than the suggested saddlepoint, which once again highlights the efficiency of the latter.



(a) Empirical coverage probability (nominal 95% CI).

(b) Average CI width (nominal 95% CI).

Figure 3. Confidence interval performance comparison (nominal 95% CI) across representative scenarios. The empirical coverage probabilities are shown in Panel (a), which shows the similarity of the 95% confidence interval of each method to the true treatment effect (Figure 3(a)). The average width of these confidence intervals is shown in panel (b) and this is the precision of estimation (Figure 3(b)). Both panels point out together the high accuracy (through coverage) and high precision (through narrower widths) of the proposed saddlepoint method over the asymptotic normal method.

5.1.2. Discussion of simulation results

The simulation findings have solid arguments to support that the proposed saddlepoint approximation furnishes an effective and precise inferential framework of linear rank tests when the randomized block urn design is considered. The main conclusions are as follows:

- As detailed in Table 2 and Figure 1, Type-I error rates in the saddlepoint method were seen to be kept at or near the nominal 0.05 level in all cases. Its simulation was almost as fast as the computationally-intensive simulated mid-p benchmark, which showed its validity in a finite sample. By contrast, the traditional asymptotic normal approximation very often had substantially inflated Type-I error rates which are usually much greater than the nominal one, reflecting the inability to capture properly the tangled dependencies that are caused by the urn mechanism and clustering.
- Table 3 and Figure 2 demonstrate the better performance of the saddlepoint estimation. In all the considered cases, the proposed method had more power than the asymptotic normal method. As an example, in Scenario 4 the saddlepoint method had a power of 76.5% as compared to 65.2% in the asymptotic approach. This is a significant advancement in the capability of identifying the real treatment effects in clinical research.
- The empirical analysis of confidence intervals (Table 4 and Figure 3) reinforces these conclusions. The intervals obtained using the saddlepoint, always attained coverage probabilities close to the nominal 95% coverage, and the asymptotic normal intervals often had undercoverage. Moreover, the saddlepoint intervals were, on the average, narrower and they gave more accurate estimation of the treatment effects without reducing reliability.
- This is because the superiority of the proposed method is based on the fact that it has the capacity of modeling the actual finite-sample distribution of the test statistic. Through the conditional CGF the saddlepoint approximation directly takes into consideration the complex stochastic structure generated by adaptive allocation and inter-cluster correlation. This is in opposition to large-sample procedures that use central limit theorem that are slow to converge in such complicated environments.
- Even though saddlepoint approximations are computationally more expensive than standard normal approximations, they are as accurate as standard permutation tests but at a small fraction of the computational cost. They are especially useful with large or complicated designs of trials in which full permutation testing is not possible. The gains in error control and precision that are observed explain the moderate rise in computational effort.
- We assessed the efficiency of the proposed method relative to that of the benchmark simulated mid-p intervals by comparing the average widths of the saddlepoint confidence intervals to the benchmark simulated mid-p intervals. The widths ratio was always very near to 1.00 (maximum variation 0.99 to 1.02), which means that the efficiency of the exact test is not lost using the saddlepoint method. Notably, such efficiency did not decrease even in the situation when intra-cluster correlation was high (e.g., Scenario 6, $\rho = 0.7$). This substantiates that the saddlepoint framework is able to adapt to the decrease in effective sample size due to clustering, a property that the asymptotic framework does not have.

5.2. *Illustrative applications*

In order to supplement the simulation study, we used the proposed method on two specific datasets created to simulate clinical trials structures in the real world. Such examples demonstrate how the method works in single-realization conditions of clustering and adaptive allocation.

5.2.1. Application 1: Glaucoma treatment (paired data)

We have a dataset representing a multi-center glaucoma trial involving paired eye data, see [31,32]. The design consisted of $K = 15$ blocks (clinics). To reflect realistic variability, the number of patients per clinic was drawn from a discrete uniform distribution $U(2, 4)$, with each patient contributing two subunits (eyes). This structure matches simulation scenario 5 (Table 1).

- **Randomization:** Treatments were assigned using a block urn design with parameters $\gamma = 1.0$ and $\alpha = 1.0$.
- **Outcome:** Continuous intraocular pressure (IOP) reduction, simulated with an intra-cluster correlation of $\rho = 0.5$ and a treatment effect of $\Delta = 1.0$.
- **Analysis:** We applied the Mann-Whitney U-test. The p-values and 95% confidence intervals were computed using the proposed saddlepoint method, the asymptotic normal approximation, and the simulated mid-p benchmark (10^6 replicates).

Table 5 gives the results of the analysis. The saddlepoint approximation gave a p-value of 0.012, which is quite close to the simulated mid-p value (0.011). The saddlepoint inversion of the 95% confidence interval was $[2.31, 5.67]$. In comparison, the asymptotic normal approximation produced a wider interval $[1.52, 6.84]$ and a p-value of 0.038.

Table 5. Hypothetical results: Glaucoma treatment study (Mann-Whitney U-test).

Method	***p*-value**	**95% Confidence Interval**		**Coverage Prob.**	**CI Width**
		Lower Bound	**Upper Bound**		
Proposed saddlepoint (SP)	**0.012**	2.31	5.67	0.949	3.36
Asymptotic normal (AN)	0.038	1.52	6.84	0.910	5.32
Simulated mid-p (Benchmark)	**0.011**	2.30	5.68	0.951	3.38

*Note: CI values are for a treatment effect parameter derived from the test, e.g., location shift.

5.2.2. Application 2: Multicenter oncology trial (survival data)

We simulated three different datasets that modeled multicenter oncology trials whose endpoint was the time-to-event [30].

- **Design:** Each dataset simulated a trial with $K = 20$ sites (blocks).
- **Correlation:** Patient outcomes within sites were correlated (frailty variances corresponding to $\rho \in \{0.3, 0.5, 0.7\}$).
- **Randomization:** Adaptive urn parameters were given in the scenario of 4, 6 and 7 where patients were assigned sequentially.

Tables 6–8 describe the findings on the Gehan-Wilcoxon, log-rank and Prentice-Wilcoxon tests, respectively. The saddlepoint confidence intervals in each of the three cases gave narrower intervals than the asymptotic normal intervals and had coverage probabilities close to the nominal 95% level compared to the benchmark.

Table 6. Hypothetical results: Multicenter oncology study (Gehan-Wilcoxon test).

Method	***p*-value**	**95% Confidence Interval**		**Coverage Prob.**	**CI Width**
		Lower Bound	**Upper Bound**		
Proposed saddlepoint (SP)	**0.007**	0.450	0.980	0.949	0.530
Asymptotic normal (AN)	0.021	0.380	1.100	0.908	0.720
Simulated mid-*p* (Benchmark)	**0.006**	0.430	0.960	0.943	0.530

*Note: CI values are for a treatment effect parameter derived from the test, e.g., log-hazard ratio.

Table 7. Hypothetical results: Multicenter oncology study (log-rank test).

Method	***p*-value**	**95% Confidence Interval**		**Coverage Prob.**	**CI Width**
		Lower Bound	**Upper Bound**		
Proposed saddlepoint (SP)	**0.018**	0.125	0.975	0.948	0.850
Asymptotic normal (AN)	0.045	0.040	1.060	0.902	1.020
Simulated mid-*p* (Benchmark)	**0.017**	0.135	0.965	0.947	0.830

*Note: CI values represent a treatment effect parameter, e.g., log-hazard ratio.

Table 8. Hypothetical results: Multicenter oncology study (Prentice-Wilcoxon test).

Method	***p*-value**	**95% Confidence Interval**		**Coverage Prob.**	**CI Width**
		Lower Bound	**Upper Bound**		
Proposed saddlepoint (SP)	**0.010**	0.200	1.100	0.950	0.900
Asymptotic normal (AN)	0.030	0.075	1.225	0.915	1.150
Simulated mid-*p* (Benchmark)	**0.009**	0.210	1.090	0.945	0.880

*Note: CI values represent a treatment effect parameter, e.g., log-hazard ratio.

6. Conclusions

The paper has developed a saddlepoint approximation framework that is used to make precise statistical inferences in clinical trials with clustered outcomes and adaptive randomization, which is the randomized block urn design. We have derived the conditional cumulant generating function of the rank test statistic, which is an explicit treatment of the two complexities of intra-cluster correlation and dynamic treatment allocation.

Our extensive simulation experiments and examples are supporting the empirical evidence of the clear benefits of this methodology compared to traditional asymptotic methods. The suggested framework always provides the mid-p-values that are in close proximity of the precise permutation values, which guarantees strong statistical support even in small sample sizes. Also, it generates accurate confidence intervals which retain nominal coverage rates but have smaller average widths thus provide more informative treatment effects estimates. Regarding hypothesis testing, the approach has been shown to be better at controlling Type-I error and has higher statistical power, which is needed to correct the situation with traditional normal approximations of the test, which is inflated and less sensitive.

Those results point to the shortcomings of traditional asymptotic approaches in that regard, which are often not able to reveal the particular finite-sample distribution caused by the design. The saddlepoint approximation that includes the higher-order cumulants of the test statistic provides a tool to overcome these shortcomings by providing a rigorous and reliable inferential tool. It is a viable methodology that offers a workable solution to researchers who are undertaking analysis of complex clinical trials in which the determination of patient safety and efficacy of treatment depends on accurate statistical summaries. Future studies will consider this framework to other categories of covariate-adaptive designs and more general hierarchical data structures to further expand the range of biostatistics applications of saddlepoint methods.

Author contributions

Haidy A. Newer: Conceptualization, methodology, theoretical development, formal analysis, software, simulation study, data analysis, visualization, data curation, writing–original draft, writing–review and editing; Bader S. Alanazi: Validation, data curation, visualization, review and funding acquisition. All authors have read and agreed to the published version of the manuscript.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors also acknowledge the Deanship of Scientific Research of Northern Border University, Arar, KSA, as the body that financed this research work with the project number “NBU-FPEJ-2026-1253-01”.

Conflict of interest

The authors declare no conflict of interest.

References

1. E. F. Abd-Elfattah, Saddlepoint p-values and confidence intervals for the class of linear rank tests for censored data under generalized randomized block design, *Computation. Stat.*, **30** (2015), 593–604. https://doi.org/10.1007/s00180-014-0551-9

2. J. G. Booth, R. W. Butler, Randomization distributions and saddlepoint approximations in generalized linear models, *Biometrika*, **77** (1990), 787–796. https://doi.org/10.1093/biomet/77.4.787

3. A. Burton, D. G. Altman, P. Royston, D. J. Holder, The design of simulation studies in medical statistics, *Stat. Med.*, **25** (2006), 4277–4292. https://doi.org/10.1002/sim.2673

4. H. Daniels, Saddlepoint approximation in statistics, *Ann. Math. Stat.*, **25** (1954), 631–650. https://doi.org/10.1214/aoms/1177728652

5. A. C. Davison, D. H. Hinkley, Saddlepoint approximations in resampling method, *Biometrika*, **75** (1988), 417–431. https://doi.org/10.1093/biomet/75.3.417

6. M. W. Fagerland, S. Lydersen, P. Laake, The McNemar test for binary matched-pairs data: Mid-p and asymptotic are better than exact conditional, *BMC Med. Res. Methodol.*, **13** (2013), 1–8. https://doi.org/10.1186/1471-2288-13-91

7. B. Friedman, A simple urn model, *Commun. Pur. Appl. Math.*, **2** (1949), 59–70. https://doi.org/10.1002/cpa.3160020103

8. E. A. Gehan, A generalized Wilcoxon test for comparing arbitrarily single censored samples, *Biometrika*, **52** (1965), 203–223. https://doi.org/10.2307/2333825

9. R. J. Gray, A class of K-sample tests for comparing the cumulative incidence of a competing risk, *Ann. Stat.*, **16** (1988), 1141–1154. https://doi.org/10.1214/aos/1176350951

10. D. D. Hanagal, A. Pandey, Gamma frailty models for bivariate survival data, *J. Stat. Comput. Sim*, **85** (2015), 3172–3189. https://doi.org/10.1080/00949655.2014.958086

11. Z. Hu, X. Du, Saddlepoint approximation reliability method for quadratic functions in normal variables, *Struct. Saf.*, **71** (2018), 24–32. https://doi.org/10.1016/j.strusafe.2017.11.001

12. J. Jeong, S. Jung, Rank tests for clustered survival data when dependent subunits are randomized, *Stat. Med.*, **25** (2006), 361–373. https://doi.org/10.1002/sim.2218

13. E. L. Kaplan, P. Meier, Nonparametric estimator from incomplete observations, *J. Am. Stat. Assoc.*, **53** (1958), 457–481. https://doi.org/10.2307/2281868

14. D. Kim, A. Agresti, Improved exact inference about conditional association in three-way contingency tables, *J. Am. Stat. Assoc.*, **90** (1995), 632–639. https://doi.org/10.1080/01621459.1995.10476557

15. K. Y. Liang, S. L. Zeger, B. Qaqish, Multivariate regression analyses for categorical data, *J. Roy. Stat. Soc. B*, **54** (1992), 3–24. https://doi.org/10.1111/j.2517-6161.1992.tb01862.x

16. D. Meng, Y. Guo, Y. Xu, S. Yang, Y. Guo, L. Pan, Saddlepoint approximation method in reliability analysis: A review, *CMES-Comp. Model. Eng.*, **139** (2024), 2329–2359. https://doi.org/10.32604/cmes.2024.047507

17. D. Meng, S. Yang, T. Lin, J. Wang, H. Yang, Z. Lv, RBMDO using Gaussian mixture model-based second-order mean-value saddlepoint approximation, *CMES-Comp. Model. Eng.*, **132** (2022), 553–568. https://doi.org/10.32604/cmes.2022.020756

18. J. V. Monaco, M. Gorfine, L. Hsu, General semiparametric shared frailty model: Estimation and simulation with frailtySurv, *J. Stat. Softw.*, **86** (2018), 1–42. https://doi.org/10.18637/jss.v086.i04

19. H. A. Newer, Saddle-point p-values and confidence intervals based on log-rank tests when dependent subunits of clustered survival data are randomized by random allocation design, *Commun. Stat.-Theor. M.*, **52** (2023), 4072–4082. https://doi.org/10.1080/03610926.2021.1986532

20. H. A. Newer, A. Abd-El-Monem, Saddlepoint approximation for weighted log-rank tests based on block truncated binomial design, *J. Biopharm. Stat.*, **33** (2023), 210–219. https://doi.org/10.1080/10543406.2022.2108825

21. H. A. Newer, The weighted log-rank tests based on stratified clustered survival data: Saddle-point p-values and confidence intervals, *J. Biopharm. Stat.*, **33** (2023), 544–554. https://doi.org/10.1080/10543406.2022.2162070

22. H. A. Newer, Saddlepoint approximation p-values of weighted log-rank tests based on censored clustered data under block Efron's biased-coin design, *Stat. Methods Med. Res.*, 2023, 1–9. https://doi.org/10.1177/09622802221143498

23. H. A. Newer, P-values and confidence intervals for weighted log-rank tests under truncated binomial design based on clustered medical data, *J. Biopharm. Stat.*, **35** (2025), 473–484. https://doi.org/10.1080/10543406.2024.2341676

24. H. A. Newer, Accurate and efficient P-values for rank-based independence tests with clustered data using a saddlepoint approximation, *Sci. Rep.*, **15** (2025), 41816. https://doi.org/10.1038/s41598-025-26728-0

25. H. A. Newer, A saddlepoint framework for accurate inference in multicenter clinical trials with imbalanced clusters, *Stat. Med.*, **45** (2026), e70408. https://doi.org/10.1002/sim.70408

26. D. Oakes, On the intransitivity of the win ratio, *Stat. Probabil. Lett.*, **216** (2025), 110267. https://doi.org/10.1016/j.spl.2024.110267

27. S. J. Pocock, C. A. Ariti, T. J. Collier, D. Wang, The win ratio: A new approach to the analysis of composite endpoints in clinical trials based on clinical priorities, *Eur. Heart J.*, **33** (2012), 176–182. https://doi.org/10.1093/eurheartj/ehr352

28. R. L. Prentice, Linear rank tests with right censored data, *Biometrika*, **65** (1978), 167–179. https://doi.org/10.1093/biomet/65.1.167

29. J. Robinson, Saddlepoint approximations for permutation tests and confidence intervals, *J. Roy. Stat. Soc.*, **44** (1982), 91–101. https://doi.org/10.1111/j.2517-6161.1982.tb01191.x

30. W. F. Rosenberger, J. M. Lachin, *Randomization in clinical trials: Theory and practice*, Wiley, 2015. https://doi.org/10.1002/9781118742112

31. B. Rosner, Multivariate methods in ophthalmology with application to other paired data situations, *Biometrics*, **40** (1984), 1025–1035. https://doi.org/10.2307/2531153

32. B. Rosner, D. Grove, Use of the Mann–Whitney U-test for clustered data, *Stat. Med.*, **18** (1999), 1387–1400. https://doi.org/10.1002/(SICI)1097-0258(19990615)18:11%3C1387::AID-SIM126%3E3.0.CO;2-V

33. I. Skovgaard, Saddlepoint expansions for conditional distributions, *J. Appl. Probab.*, **24** (1987), 875–887. https://doi.org/10.2307/3214212

34. M. E. Terry, Some rank order tests which are most powerful against specific parametric alternatives, *Ann. Math. Stat.*, (1952), 346–366. Available from: `https://www.jstor.org/stable/2236679`.

35. P. H. Van Elteren, On the combination of independent two sample tests of Wilcoxon, *Bull. Inst. Internat. Statist.*, **37** (1960), 351–361.

36. L. J. Wei, *The design for the control of selection bias* (technical report), University of South Carolina-Columbia, Department of Mathematics and Computer Science, 1975.

37. L. J. Wei, A class of designs for sequential clinical trials, *J. Am. Stat. Assoc.*, **72** (1977), 382–386. https://doi.org/10.1080/01621459.1977.10481005

38. L. J. Wei, The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis, *Stat. Med.*, **11** (1992), 1871–1879. https://doi.org/10.1002/sim.4780111409

39. Y. Zhang, W. F. Rosenberger, R. T. Smythe, Sequential monitoring of randomization tests: Stratified randomization, *Biometrics*, **63** (2017), 865–872. https://doi.org/10.1111/j.1541-0420.2006.00735.x

40. W. Zhao, Y. Weng, Block urn design: A new randomization algorithm for sequential trials with two or more treatments and balanced or unbalanced allocation, *Contemp. Clin. Trials*, **32** (2011), 953–961. https://doi.org/10.1016/j.cct.2011.08.004